Title
Structure or Noise?

Authors
Still, Susanne
Crutchfield, James P

Publication Date
2007-08-04

Peer reviewed

Santa Fe Institute Working Paper 07-08-020
arxiv.org:0708.0654 [physics.gen-ph]

Structure or Noise?

Susanne Still[1,*] and James P. Crutchfield[2,†]

[1]*Information & Computer Sciences, University of Hawaii at Manoa, Honolulu, HI 96822*
[2]*Complexity Sciences Center & Physics Department, University of California Davis, One Shields Avenue, Davis, CA 95616*

(Dated: October 22, 2018)

We show how rate-distortion theory provides a mechanism for automated theory building by naturally distinguishing between regularity and randomness. We start from the simple principle that model variables should, as much as possible, render the future and past conditionally independent. From this, we construct an objective function for model making whose extrema embody the trade-off between a model's structural complexity and its predictive power. The solutions correspond to a hierarchy of models that, at each level of complexity, achieve optimal predictive power at minimal cost. In the limit of maximal prediction the resulting optimal model identifies a process's intrinsic organization by extracting the underlying causal states. In this limit, the model's complexity is given by the statistical complexity, which is known to be minimal for achieving maximum prediction. Examples show how theory building can profit from analyzing a process's *causal compressibility*, which is reflected in the optimal models' rate-distortion curve—the process's characteristic for optimally balancing structure and noise at different levels of representation.

PACS numbers: 02.50.-r 89.70.+c 05.45.Tp 02.50.Ey

I. INTRODUCTION

Progress in science is often driven by the discovery of novel patterns. Historically, physics has relied on the creative mind of the theorist to articulate mathematical models that capture nature's regularities in physical principles and laws. But the last decade has witnessed a new era in collecting truly vast data sets. Examples include contemporary experiments in particle physics [1] and astronomy [2], but range to genomics, automated language translation [3], and web social organization [4]. In all these, the volume of data far exceeds what any human can analyze directly by hand.

This presents a new challenge—automated pattern discovery and model building. A principled understanding of model making is critical to provide theoretical guidance for developing automated procedures. In this Letter, we show how basic information-theoretic optimality criteria provide a method for automatically constructing a hierarchy of models that achieve different degrees of abstraction. Importantly, we show that in appropriate limits the method recovers a process's causal organization. Without this connection, it would be only another approach to statistical inference, with its own ad hoc assumptions about the character of natural pattern.

Our starting point is the observation that natural systems store, process, and produce information—they compute intrinsically [5, 6, 7]. Theory building, then, faces the challenge of extracting from that information the structures underling its generation. Any physical theory delineates mechanism from randomness by identifying what part of an observed phenomenon is due to the underlying process's structure and what is irrelevant. Irrelevant parts are considered noise and typically modeled probabilistically. Successful theory building therefore depends centrally on deciding what is structure and what is noise; often, an implicit distinction.

What constitutes a good theory, though? Which information is relevant? One can answer this question for time series prediction: Information about the future of the time series is relevant. Beyond forecasting, though, models are often put to the test by assessing how well they predict new data and, hence, it is of general importance that a model capture information which aids prediction. Typically, there are many models that explain a given data set, and between two models that are equally predictive, one favors the simpler, smaller, less structurally complex model [8, 9]. However, a more complex model can achieve smaller prediction error than a less complex model. The trade-off between model complexity and prediction error is tantamount to finding a distinction between causal structure and noise.

The trade-off between assigning a causal mechanism to the occurrence of an event or explaining the event as being merely random has a long history, but how one implements the trade-off is still a very active topic. Nonlinear time series analysis [10, 11, 12], to take one example, attempts to account for long-range correlations produced by nonlinear dynamical systems—correlations not adequately modeled by assumptions such as linearity and independent, identically distributed (i.i.d.) data. Success in this endeavor requires directly addressing the notion of structure and pattern [10, 13].

Examination of the essential goals of prediction led to a principled definition of structure that captures a dynamical system's causal organization in part by discovering the underlying *causal states* [5, 6, 7]. In *computational*

*Electronic address: sstill@hawaii.edu
†Electronic address: chaos@cse.ucdavis.edu

mechanics a process $\mathrm{P}(\overleftarrow{X}, \overrightarrow{X})$ is viewed as a communication channel [7, 14]: it transmits information from the *past* $\overleftarrow{X} = \ldots X_{-3}X_{-2}X_{-1}$ to the *future* $\overrightarrow{X} = X_0X_1X_2\ldots$ by storing it in the present. For the purpose of forecasting the future two different pasts, say $\overleftarrow{x}$ and $\overleftarrow{x}'$, are equivalent if they result in the same prediction [5]. In general this prediction is probabilistic, given by the conditional future distribution $\mathrm{P}(\overrightarrow{X} \,|\overleftarrow{x})$. The resulting equivalence relation $\overleftarrow{x} \sim \overleftarrow{x}'$ groups all histories that give rise to the same conditional future distribution:

$$\epsilon(\overleftarrow{x}) = \{\overleftarrow{x}' : \Pr(\overrightarrow{X} \,|\overleftarrow{x}) = \Pr(\overrightarrow{X} \,|\overleftarrow{x}')\}. \tag{1}$$

The resulting partition of the space $\overleftarrow{\mathbf{X}}$ of pasts defines the process's *causal states* $\boldsymbol{\mathcal{S}} = \mathrm{P}(\overleftarrow{X}, \overrightarrow{X})/\sim$.

The causal states constitute a model that is maximally predictive by means of capturing all the information that the past of a time series contains about the future. As a result, knowing the causal state renders past and future conditionally independent, a property we call *causal shielding*, because the causal states have the Markovian property that they *shield* past and future [7]:

$$\mathrm{P}(\overleftarrow{X}, \overrightarrow{X} \,|\mathcal{S}) = \mathrm{P}(\overleftarrow{X} \,|\mathcal{S})\mathrm{P}(\overrightarrow{X} \,|\mathcal{S}), \tag{2}$$

where $\mathcal{S} \in \boldsymbol{\mathcal{S}}$. This is related to the fact that the causal-state partition is optimally predictive. To see this, note that Eq. (2) implies $\mathrm{P}(\overrightarrow{X} \mid \overleftarrow{X}, \mathcal{S}) = \mathrm{P}(\overrightarrow{X} \,|\mathcal{S})$. Furthermore, note that, by definition, for *any* partition $\boldsymbol{\mathcal{R}}$ of $\overleftarrow{\mathbf{X}}$ with states $\mathcal{R}$, when the past is known, then the future distribution is not altered by the history-space partitioning:

$$\mathrm{P}(\overrightarrow{X} \mid \overleftarrow{X}, \mathcal{R}) = \mathrm{P}(\overrightarrow{X} \mid \overleftarrow{X}) \ . \tag{3}$$

This implies for the causal states that $\mathrm{P}(\overrightarrow{X} \mid \overleftarrow{X}, \mathcal{S}) = \mathrm{P}(\overrightarrow{X} \mid \overleftarrow{X})$ and thus $\mathrm{P}(\overrightarrow{X} \,|\mathcal{S}) = \mathrm{P}(\overrightarrow{X} \mid \overleftarrow{X})$. Therefore, causal shielding is equivalent to the fact [7] that the causal states capture *all* of the information $I[\overleftarrow{X}; \overrightarrow{X}]$ that is shared between past and future: $I[\mathcal{S}; \overrightarrow{X}] = I[\overleftarrow{X}; \overrightarrow{X}]$, the process's *excess entropy* $\mathbf{E}$ or *predictive information* [15, 16, and references therein].

The causal states are *unique and minimal sufficient statistics* for time series prediction, capturing all of a process's predictive information at maximum efficiency [7]. The causal-state partition has the smallest *statistical complexity*, $C_\mu := H(\mathcal{S}) \leq H[\widehat{\mathcal{R}}]$, compared to all other equally predictive partitions $\widehat{\boldsymbol{\mathcal{R}}}$. C_μ measures the minimal amount of information that must be stored in order to communicate all of the excess entropy from the past to the future. Briefly stated, the causal states serve as the basis against which alternative models should be compared.

II. CONSTRUCTING CAUSAL MODELS USING RATE-DISTORTION THEORY

There are many scenarios in which one does not need to or explicitly does not want to capture *all* of the predictive information. How can we approximate the causal states in a controlled way?

In this Letter, we show how to systematically construct smaller models, which are necessarily less predictive, but which are optimal in the sense that they capture, at a fixed model complexity, the maximum possible amount of predictive information. Importantly, in the limit that removes the constraint on model complexity, our method retrieves the exact causal-state partition.

Appealing to information theory again, we frame this in terms of communicating a model over a channel with limited capacity. Rate-distortion theory [17] provides a principled way to find a lossy compression of an information source such that the resulting code is minimal at fixed fidelity to the original signal.

The compressed representation, denote it $\boldsymbol{\mathcal{R}}$, is in general specified by a *probabilistic* map $\mathrm{P}(\mathcal{R}|\overleftarrow{x})$ from the input message, here the past $\overleftarrow{x}$, to code words, here the model's states $\mathcal{R}$ with values $\rho \in \boldsymbol{\mathcal{R}}$. In contrast, Eq. (1) specifies models that are described by a deterministic map from histories to states: The causal states $\sigma \in \boldsymbol{\mathcal{S}}$ induce a deterministic partition of $\overleftarrow{\mathbf{X}}$ [7], as one can show that $\mathrm{P}(\sigma|\overleftarrow{x}) = \delta_{\sigma,\epsilon(\overleftarrow{x})}$. The mapping $\mathrm{P}(\mathcal{R}|\overleftarrow{x})$ specifies a model, and the *coding rate* $I[\overleftarrow{X}; \mathcal{R}]$ measures its complexity, which in turn is related to its statistical complexity via $I[\overleftarrow{X}; \mathcal{R}] = H[\mathcal{R}] - H[\mathcal{R}|\overleftarrow{X}] = C_\mu(\mathcal{R}) - H[\mathcal{R}|\overleftarrow{X}]$. For deterministic partitions the statistical complexity and the coding rate are equal, because then $H[\mathcal{R}|\overleftarrow{X}] = 0$. However, for more general, nondeterministic partitions, $H[\mathcal{R}|\overleftarrow{X}] \neq 0$, meaning that the probabilistic nature of the mapping curtails some of the model's complexity, and the coding rate $I[\overleftarrow{X}; \mathcal{R}]$ captures this.

To illustrate this point, consider the extreme of uniform assignments: $\mathrm{P}(\mathcal{R}|\overleftarrow{x}) = 1/c$, for any given $\overleftarrow{x}$, where $c = |\boldsymbol{\mathcal{R}}|$. In this case, even if there are many states—large statistical complexity $H[\mathcal{R}] = \log_2(c)$—they are indistinguishable: $\mathrm{P}(\overrightarrow{x} \,|\mathcal{R}) = \langle \mathrm{P}(\overrightarrow{x} \,|\overleftarrow{x})\rangle_{\mathrm{P}(\overleftarrow{x})}$, for all $\mathcal{R}$, due to the large uncertainty about the state, given the past. This is reflected in $H[\mathcal{R}|\overleftarrow{X}] = \log_2(c)$. In effect, the model has only one state (the average $\langle \mathrm{P}(\overrightarrow{x} \,|\overleftarrow{x})\rangle_{\mathrm{P}(\overleftarrow{x})}$) and its statistical complexity vanishes, which is reflected in the coding rate: $I[\overleftarrow{X}; \mathcal{R}] = 0$.

Rate-distortion theory allows us to back away from the best (causal-state) representation toward less complex models by controlling the coding rate: Simpler models are distinguished from more complex ones by the fact that they can be transmitted more concisely. However, less complex models are also associated with a larger error. Rate-distortion theory quantifies the loss by a *distor-*

tion function $d(\overleftarrow{x};\rho)$. The coding rate is then minimized [14] over the assignments $\mathrm{P}(\mathcal{R}|\ \overleftarrow{X})$ at fixed average distortion $D[\overleftarrow{X};\mathcal{R}] = \left\langle d(\overleftarrow{x};\rho)\right\rangle_{\mathrm{P}(\overleftarrow{x},\rho)}$.

In building predictive models, the loss should be measured by how much the resulting models deviate from accurate prediction. We take the shielding property, Eq. (2), of the causal-state partition as the goal for any predictive model. This condition is equivalent to the statement that the excess entropy *conditioned on the model states* $\mathcal{R}$:

$$I[\overleftarrow{X};\overrightarrow{X}\,|\mathcal{R}] = \left\langle\left\langle \log\left[\frac{\mathrm{P}(\overleftarrow{x},\overrightarrow{x}\,|\rho)}{\mathrm{P}(\overrightarrow{x}\,|\rho)\mathrm{P}(\overleftarrow{x}|\rho)}\right]\right\rangle_{\mathrm{P}(\overrightarrow{x}|\overleftarrow{x})}\right\rangle_{\mathrm{P}(\overleftarrow{x},\rho)} \quad (4)$$

vanishes for the causal-state partition: $I[\overleftarrow{X};\overrightarrow{X}\,|\mathcal{S}] = 0$. This gives us our distortion measure:

$$d(\overleftarrow{x};\rho) = \left\langle \log\left[\frac{\mathrm{P}(\overleftarrow{x},\overrightarrow{x}\,|\rho)}{\mathrm{P}(\overleftarrow{x}|\rho)\mathrm{P}(\overrightarrow{x}\,|\rho)}\right]\right\rangle_{\mathrm{P}(\overrightarrow{x}|\overleftarrow{x})} . \quad (5)$$

From Eq. (3) this is the same as the relative entropy between the conditional future distributions given the past and those given the model states ρ:

$$\mathcal{D}\left(\mathrm{P}(\overrightarrow{x}\,|\overleftarrow{x})||\mathrm{P}(\overrightarrow{x}\,|\rho)\right) = \left\langle \log\left[\frac{\mathrm{P}(\overrightarrow{x}\,|\overleftarrow{x})}{\mathrm{P}(\overrightarrow{x}\,|\rho)}\right]\right\rangle_{\mathrm{P}(\overrightarrow{x}|\overleftarrow{x})} \quad (6)$$

Altogether, we solve the constrained optimization problem:

$$\min_{\mathrm{P}(\mathcal{R}|\overleftarrow{X})} \left(I[\overleftarrow{X};\mathcal{R}] + \beta I[\overleftarrow{X};\overrightarrow{X}\,|\mathcal{R}]\right) , \quad (7)$$

where the Lagrange multiplier β controls the trade-off between model complexity and prediction error; i.e., the balance between structure and noise.

The conditional excess entropy of Eq. (4) is the difference between the process's excess entropy and the information $I[\mathcal{R};\overrightarrow{X}]$ that the model states contain about the future: $I[\overleftarrow{X};\overrightarrow{X}\,|\mathcal{R}] = I[\overleftarrow{X};\overrightarrow{X}] - I[\mathcal{R};\overrightarrow{X}]$, due to Eq. (3). The excess entropy $I[\overleftarrow{X};\overrightarrow{X}]$ is a property intrinsic to the process, however, and so not dependent on the model. Therefore, the optimization problem in Eq. (7) is equivalent to maximizing the information that the model states carry about the future while minimizing information kept about the past. This maps directly onto the *information bottleneck* (IB) method [18]—here the future data is IB's "relevant" quantity with respect to which the past is summarized.

In any case, the solution to the optimization principle is given by (cf. [18]):

$$\mathrm{P}_{\mathrm{opt}}(\rho|\overleftarrow{x}) = \frac{\mathrm{P}(\rho)}{Z(\overleftarrow{x},\beta)} e^{-\beta E(\rho,\overleftarrow{x})} , \quad (8)$$

where

$$E(\rho,\overleftarrow{x}) = \mathcal{D}\left(\mathrm{P}(\overrightarrow{X}\,|\overleftarrow{x})||\mathrm{P}(\overrightarrow{X}\,|\rho)\right) , \quad (9)$$

$$\mathrm{P}(\overrightarrow{X}\,|\rho) = \frac{1}{\mathrm{P}(\rho)} \sum_{\overleftarrow{x}\in\overleftarrow{\mathbf{X}}} \mathrm{P}(\overrightarrow{X}\,|\overleftarrow{x})\mathrm{P}(\rho|\overleftarrow{x})\mathrm{P}(\overleftarrow{x}) , \text{ and} \quad (10)$$

$$\mathrm{P}(\rho) = \sum_{\overleftarrow{x}\in\overleftarrow{\mathbf{X}}} \mathrm{P}(\rho|\overleftarrow{x})\mathrm{P}(\overleftarrow{x}) . \quad (11)$$

Eqs. (8)-(11) must be solved self-consistently, and this can be done numerically [18].

Eq. (8) specifies a family of models parametrized by β with the form of Gibbs distributions. Within an analogy to statistical mechanics [19], β corresponds to the inverse temperature, E is the energy, and $Z = \left\langle e^{-\beta E(\rho,\overleftarrow{x})}\right\rangle_{\mathrm{P}(\rho)}$ the partition function. Finally, note that for linear Gaussian-distributed random variables the optimal linear map can be computed analytically [20]. These results can be carried over to the temporal setting that concerns us here for linear Gaussian processes following a rate-distortion approach similar to the above [21].

III. RETRIEVING THE CAUSAL-STATE PARTITION

A key result is that these optimal solutions retrieve the causal-state partition in the limit $\beta\to\infty$, which emphasizes prediction accuracy [22, detailed proof]. To see this first note that as $\beta\to\infty$, the optimal assignment becomes deterministic $\mathrm{P}_{\mathrm{opt}}(\rho|\overleftarrow{x}) \to \delta_{\rho,\rho^*(\overleftarrow{x})}$, where the state $\rho^*(\overleftarrow{x})$ to which a past is assigned is the one minimizing energy, Eq. (9). Now, that function is zero when the future probability conditioned on the state equals the future probability conditioned on the past. This means that, in the limit, all pasts with equal conditional future probability distributions will be assigned to the same state with $\mathrm{P}(\overrightarrow{X}\,|\overleftarrow{x}) = \mathrm{P}(\overrightarrow{X}\,|\rho^*(\overleftarrow{x}))$, for all those pasts assigned to the state $\rho^*(\overleftarrow{x})$. This yields exactly the causal-state partition given by the equivalence relation that arises from Eq. (1).

Hence, one finds in this limit what we have argued is the goal of predictive modeling. Moreover, what was otherwise an ad hoc optimization method has been given a structural grounding in that it captures a process's intrinsic causal architecture. Recall that the model complexity C_μ of the causal-state partition is minimal among the optimal predictors and so not necessarily equal to the maximum value of the coding rate $I[\overleftarrow{X};\mathcal{R}] \leq H[\overleftarrow{X}]$.

IV. FINDING APPROXIMATE CAUSAL REPRESENTATIONS: CAUSAL COMPRESSIBILITY

While the causal-state partition captures *all* of the predictive information, less complex models can be con-

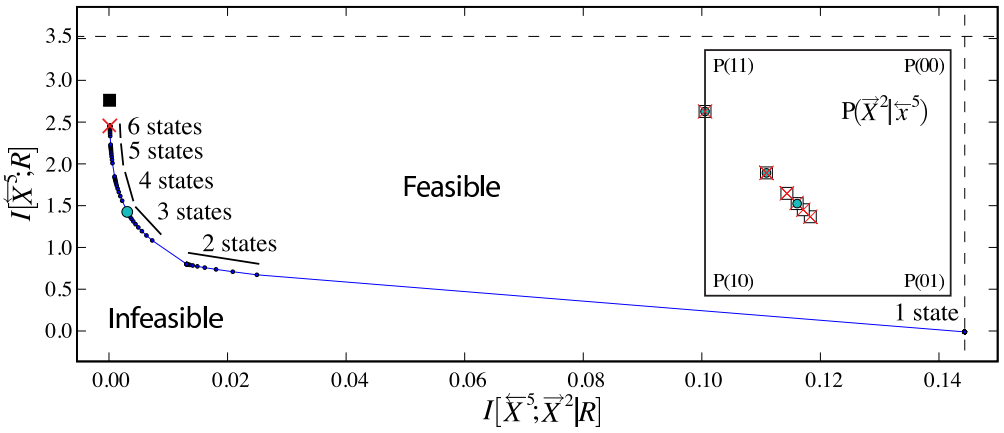

FIG. 1: Trading structure off against noise using optimal causal inference (OCI): Rate-distortion curve for the SNS process, coding rate $I[\overleftarrow{X};\mathcal{R}]$ versus distortion $I[\overleftarrow{X};\overrightarrow{X}\,|\mathcal{R}]$. Dashed lines mark maximum values: past entropy $H[\overleftarrow{X}^5]$ (horizontal) and excess entropy $I[\overleftarrow{X};\overrightarrow{X}]$ (vertical). The causal-state limit for infinite sequences is shown in the upper left (solid box). (Inset) SNS conditional future distributions $P(\overrightarrow{X}^2\,|\overleftarrow{x}^5)$: OCI six-state reconstruction (six crosses), true causal states (six boxes), and three-state approximation (three circles). Annealing rate was $\alpha = 1.1$.

structed if one allows for larger distortion—accepting less predictive power. For all models in the optimal family, Eqs. (8)-(11), the original process is mapped to the best causal-state *approximation*, at fixed model complexity. And so we refer to the resulting method as *optimal causal inference* (OCI). Several examples are studied in [22].

The nature of the trade-off embodied in Eq. (7) can be studied by evaluating the objective function at the optimum for each value of β. The shape of the resulting rate-distortion curve characterizes a process's *causal compressibility* via the interdependence between $I[\overleftarrow{X};\mathcal{R}]$ and $I[\overleftarrow{X};\overrightarrow{X}\,|\mathcal{R}]$. Since the variation of the objective function in Eq. (7) vanishes at the optimum, the curve's slope is $\delta I[\overleftarrow{X};\mathcal{R}]/\delta D[\overleftarrow{X};\mathcal{R}] = -\beta$. For a given process the rate-distortion curve determines what predictability the best model at a fixed complexity can achieve and, vice versa, how small a model can be made at fixed predictability. Below the curve lie *infeasible* causal compression codes; above are *feasible* larger models that are no more predictive that those directly on the curve. In short, the rate-distortion curve determines how to *optimally* trade structure for noise.

As an example, consider the *simple nondeterministic source* (SNS)—a hidden Markov process that specifies a binary information source with nontrivial statistical structure, including infinite-range correlations and an infinite number of causal states [29].

The SNS's rate-distortion curve, calculated for pasts of length 5 and futures of length 2 is shown in Fig. 1. We computed the curve using a deterministic annealing scheme following [19]. One starts at a high temperature (low β) and slowly cools the system, waiting for it to equilibrate—iterating the self-consistent Eqs. (8)-(11) until convergence. At that point one continues by lowering the temperature ($\beta \leftarrow \alpha\beta$) by a fixed annealing rate $\alpha > 1$ and equilibrating again. During this procedure, the number of effective states changes. Starting at high temperatures, all pasts are assigned to states that are all effectively the same state, as their predictions are equal. States are allowed to split at each temperature. One observes the proliferation of more and more states as the temperature is lowered, until the causal states emerge in the zero-temperature limit.

For the SNS the causal states for past and future strings of *finite* length are recovered by OCI (cross in upper left). For a comparison, there we also show the *causal-state limit*, which is calculated analytically for *infinite* pasts and futures (solid box).

The curve drops rapidly away from the finite causal-state model with six effective states, indicating that there is little predictive cost in using significantly smaller models with successively fewer effective states. The curve then levels out below three states: smaller models incur a substantial increase in distortion (loss in predictability) while little is gained in terms of compression. Quantitatively, specifying the best four-state model (at $I[\overleftarrow{X};\mathcal{R}] = 1.92$ bits) leads to 0.5% distortion, capturing 99.5% the SNS's excess entropy. The distortion increases to 2% for three states (1.43 bits), 9% for two states (0.81 bits), and 100% for a single state (0 bits). Overall, the three-state model lies near a knee in the rate-distortion curve and this suggests that it is a good compromise between model complexity and predictability.

The inset in Fig. 1 shows the reconstructed conditional future distributions for the optimal three-state and six-state models in the simplex $P(\overrightarrow{X}^2\,|\overleftarrow{x}^5)$. The six-state model (crosses) reconstructs the true causal-state conditional future distributions (boxes), calculated from analytically known finite-sequence causal states. The figure illustrates why the three-state model (circles) is a good compromise: two of the three-state model's conditional future distributions capture the two more-distinct SNS

conditional future distributions, and its third one summarizes the remaining, less different, SNS conditional future distributions.

With its intricate causal structure and nontrivial causal compressibility properties the SNS process is typical of stochastic processes. Other frequently studied processes are not, however. Two classes are of particular interest due to their widespread use. On one extreme of randomness are the i.i.d. processes alluded to in the introduction, such as the biased coin—by definition, a completely random and unstructured source. For all i.i.d. processes the rate-distortion curve collapses to a single point at $(0,0)$, indicating that they are wholly unpredictable and causally incompressible. This is easily seen by noting first that for i.i.d. processes the excess entropy $I[\overleftarrow{X};\overrightarrow{X}]$ vanishes, since $\mathrm{P}(\overrightarrow{x}\ |\overleftarrow{x}) = \mathrm{P}(\overrightarrow{x})$. Therefore, $I[\overleftarrow{X};\overrightarrow{X}\ |\mathcal{R}] \leq I[\overleftarrow{X};\overrightarrow{X}] = 0$ vanishes, too. Second, the energy function $E(\rho,\overleftarrow{x})$ in the optimal assignments, Eq. (9), vanishes, since $\mathrm{P}(\overrightarrow{x}\ |\rho) = \left\langle \mathrm{P}(\overrightarrow{x}\ |\overleftarrow{x}) \right\rangle_{\mathrm{P}(\overleftarrow{X}|\rho)} = \mathrm{P}(\overrightarrow{x})$. The optimal assignment given by Eq. (8) is therefore the uniform distribution and $I[\overleftarrow{X};\mathcal{R}]|_{\mathrm{P}_{\mathrm{opt}}(\rho|\overleftarrow{x})} = 0$. (See Fig. 2.)

At the other extreme are the *predictively reversible* processes for which

$$\mathrm{P}(\overrightarrow{x}\ |\overleftarrow{x}) = \delta_{\overrightarrow{x},f(\overleftarrow{x})}, \tag{12}$$

where f is *invertible*, such as periodic processes. These processes have a rate-distortion curve that is a straight line, the negative diagonal. Note that $\mathrm{P}(\overrightarrow{x}\ |\rho) = \mathrm{P}(f^{-1}(\overleftarrow{x})|\rho) = \mathrm{P}(\overleftarrow{x}|\rho)$ and, therefore, $I[\overleftarrow{X};\overrightarrow{X}\ |\mathcal{R}] = I[\overleftarrow{X};\overrightarrow{X}] - I[\overleftarrow{X};\mathcal{R}]$. The variational principle now reads $\delta(1-\beta)I[\overleftarrow{X};\mathcal{R}] = 0$, which implies that $\beta = 1$. For these processes, the rate-distortion curve is the diagonal that runs from $[0, C_\mu]$ (causal-state limit) to $[\mathbf{E}, 0]$, where $\mathbf{E} = I[\overleftarrow{X};\overrightarrow{X}]$ is the excess entropy, due to Eq. (12) and invertibility. (See Fig. 2.)

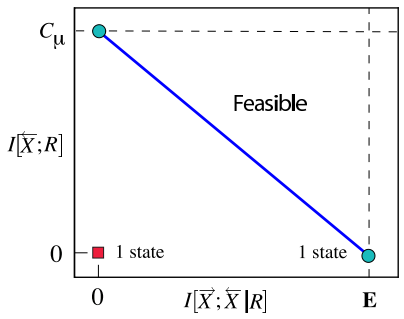

FIG. 2: Schematic illustration of the causal incompressibility of independent, identically distributed processes (square) and predictively reversible processes (straight line connecting circles).

This diagonal rate-distortion curve represents the worst possible case for causal compression. At each level, specifying the future to one bit higher accuracy costs us exactly one bit in model complexity. Processes in this class are thus not causally compressible. To be causally compressible, a process's rate-distortion curve must lie below the diagonal. The more concave the curve, the more causally compressible is the process. An extremely causally compressible process can be predicted to high accuracy with a model that can be encoded at a very low model cost. These are the processes that lie between the extremes of exact predictability and structureless randomness.

These examples show how studying the hierarchy of optimal models, and the associated rate-distortion curve, allows one to learn about the causal compressibility of the process at hand, which serves to guide where the demarcation between structure and noise should lie.

V. FINITE-SAMPLE FLUCTUATIONS

As in statistical mechanics, we assumed so far that the distribution $\mathrm{P}(\overleftarrow{X},\overrightarrow{X})$ is given. And so, the above results bear on an intrinsic distinction between structure and noise for a process, unsullied by statistical sample fluctuations.

However, when one builds a model from *finite* samples, the distributions must be estimated from the available data and so sample fluctuations must be taken into account. Intuitively, limited data size sets a bound on how much we can consider to be structure without overfitting. It turns out that using [23], the effects of finite data can be corrected, as we show in [22]. This connects the approach taken here to statistical inference and machine learning, where model complexity control is designed to avoid overfitting due to finite-sample fluctuations; cf., e.g, [24, 25, 26, 27, 28].

VI. CONCLUSION

We showed how rate-distortion theory can be employed to find optimal causal models at varying degrees of abstraction. Starting with the simple modeling principle of causal shielding, an objective function was constructed that embodied the trade-off between model complexity and predictability. Since the variational principle corresponded to a rate-distortion theory known analysis methods could be employed. Solutions to the objective function were found using an iterative algorithm, and the rate-distortion curve was computed using deterministic annealing.

For certain processes we calculated the curve analytically. These and a numerical example served to demonstrate how its shape reveals a process's causal compressibility, providing direct guidance for automated model making. In particular, we showed how a model distinguishes between what it effectively considers to be underlying structure and what is noise. Practically speaking,

natural processes that have high causal compressibility will admit particularly parsimonious theories that capture a large fraction of observed behavior.

We pointed out that OCI finds the causal-state partition exactly when the constraint on model complexity is relaxed. Then we showed how to automatically build models with varying degrees of abstraction. By focusing on the case in which limitations due to finite sampling errors are absent, we emphasized that compact representations, in and of themselves, are critical aids to scientific understanding. We pointed out, however, that finite data set size imposes a maximum level of allowable accuracy before overfitting occurs and that previous results can be used to find that demarcation line as well.

Acknowledgments

We thank Chris Ellison, on a GAANN fellowship, for programming. The CSC Network Dynamics Program funded by Intel Corporation supported this work.

[1] W. von Rueden and R. Mondardini. The Large Hadron Collider (LHC) data challenge. Technical report, IEEE Technical Committee on Scalable Computing, 2007. http://www.ieeetcsc.org/newsletters/2003-01/mondardini.html.

[2] Anonymous. LSST observatory—Baseline configuration. Technical report, LSST Corporation, Tucson, AZ, 2007. http://www.lsst.org/Science/lsst_baseline.shtml.

[3] Anonymous. NIST 2006 machine translation evaluation official results. Technical report, National Institute of Standards and Technologies, Washington, DC, 2006.

[4] M. E. J. Newman. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA*, 78(2):404–409, 2001.

[5] J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Let.*, 63:105–108, 1989.

[6] J. P. Crutchfield. The calculi of emergence: Computation, dynamics, and induction. *Physica D*, 75:11–54, 1994.

[7] J. P. Crutchfield and C. R. Shalizi. Thermodynamic depth of causal states: Objective complexity via minimal representations. *Phys. Rev. E*, 59(1):275–283, 1999.

[8] William of Ockham. *Philosophical Writings: A Selection, Translated, with an Introduction, by Philotheus Boehner, O.F.M., Late Professor of Philosophy, The Franciscan Institute*. Bobbs-Merrill, Indianapolis, 1964. first pub. various European cities, early 1300s.

[9] P. Domingos. The role of Occam's Razor in knowledge disocvery. *Data Mining and Knowledge Discovery*, 3:409–425, 1999.

[10] M. Casdagli and S. Eubank, editors. *Nonlinear Modeling*, SFI Studies in the Sciences of Complexity, Reading, Massachusetts, 1992. Addison-Wesley.

[11] J. C. Sprott. *Chaos and Time-Series Analysis*. Oxford University Press, Oxford, UK, second edition, 2003.

[12] H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis*. Cambridge University Press, Cambridge, UK, second edition, 2006.

[13] J. P. Crutchfield and B. S. McNamara. Equations of motion from a data series. *Complex Systems*, 1:417 – 452, 1987.

[14] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, second edition, 2006.

[15] J. P. Crutchfield and D. P. Feldman. Regularities unseen, randomness observed: Levels of entropy convergence. *CHAOS*, 13(1):25–54, 2003.

[16] W. Bialek, I. Nemenman, and N. Tishby. Predictability, Complexity and Learning. *Neural Computation*, 13:2409–2463, 2001.

[17] C. E. Shannon. A mathematical theory of communication. *Bell Sys. Tech. J.*, 27, 1948. Reprinted in C. E. Shannon and W. Weaver *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, 1949.

[18] N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In B. Hajek and R. S. Sreenivas, editors, *Proc. 37th Allerton Conference*, pages 368–377. University of Illinois, 1999.

[19] K. Rose. Deterministic Annealing for Clustering, Compression, Classification, Regression, and Related Optimization Problems. *Proc. IEEE*, 86(11):2210–2239, 1998.

[20] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss. Information bottleneck for Gaussian variables. *J. Machine Learning Res.*, 6:165–188, 2005.

[21] F. Creutzig. personal communication, 2008.

[22] S. Still, J. P. Crutchfield, and C. J. Ellison. Optimal causal inference. 2007. Santa Fe Institute Working Paper 2007-08-024; arxiv.org:0708.1580 [cs.IT].

[23] S. Still and W. Bialek. How many clusters? An information theoretic perspective. *Neural Computation*, 16(12):2483–2506, 2004.

[24] C. Wallace and D. Boulton. An information measure for classification. *Comput. J.*, 11:185, 1968.

[25] H. Akaike. An objective use of Bayesian models. *Ann. Inst. Statist. Math.*, 29A:9, 1977.

[26] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific, Singapore, 1989.

[27] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.

[28] D. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge, 2003.

[29] The SNS has causal states $\sigma_i, i = 0, 1, \ldots \infty$, and output-labeled transition matrices whose nonzero entries are $T^{(0)}_{i0} = \frac{1}{2}(1 - \frac{1}{i+1})$ and $T^{(1)}_{i,i+1} = \frac{1}{2}(1 + \frac{1}{i+1})$. It produces $h_\mu \approx 0.677867$ bits of information per output symbol and stores $C_\mu \approx 2.71147$ bits of historical information [6].